Empirical Evaluation of φ-Regularization in Language Models

Dmitri A. Sklyarov

Abstract

We present a comparative evaluation of language model behavior under varying strengths of phi-regularization, a novel mechanism that balances predictive power and semantic uncertainty via the phi-metric. Four levels of regularization (0.0, 0.1, 0.25, and 0.5) were tested across 15 diverse prompts spanning scientific, historical, philosophical, strategic, and creative domains.

Keywords: Large Language Models, parameter-efficient fine-tuning, φ-regularization, representation geometry, coherence, TinyLlama, FAISS.

1 Introduction

The φ-metric regulates hidden states by penalizing deviations from an optimal balance point ($\varphi_i \approx 1.0$), where tokens are neither too predictable (trivial) nor too uncertain (noisy). This work evaluates its empirical impact on model coherence, accuracy, and reliability. Code and training scripts are publicly available at: https://github.com/skl-AI/phi-metric.git.

2 Related Work

Our work builds upon several key developments in efficient language model fine-tuning and representation analysis. The Transformer architecture [7] underpins modern LLMs, including TinyLlama [9]. Efficient adaptation is enabled by LoRA [4] and its quantized extension QLoRA [2], which leverage 4-bit NF4 quantization [3]. Nearest-neighbor search for semantic analysis uses FAISS [6], and our implementation relies on the Hugging Face ecosystem [8, 5]. Our φ-metric extends recent work on geometric stability in latent spaces, as introduced in [1].

3 Methodology

3.1 Adaptation of the φ-Metric from "Mechanics of Mind"

Our work is grounded in the conceptual framework of the "Mechanics of Mind" [1], which introduces the φ-metric as a measure of balance for a unit of "pro-information":

$$\varphi_i = \frac{\text{Relations}}{\text{Copies}}.$$

Here, *Relations* denote the dependency of other system elements on the unit, while *Copies* refer to its existing realizations. The theory posits that a balanced state is achieved when $\varphi_i \approx 1.0$.

We adapt these abstract concepts to the geometric structure of Transformer hidden states, mapping them to two measurable quantities:

Uncertainty ($\mathbf{H}_i$) as "Copies". In the LLM context, the "quantity of copies" corresponds to how closely the contextualized hidden state $\mathbf{h}_i$ aligns with known semantic centers in the input embedding matrix $\mathbf{E}$. A state far from these centers indicates semantic novelty or uncertainty. We quantify this as the mean L2 distance to the K nearest neighbors in $\mathbf{E}$:

$$\mathbf{H}_i = \text{mean}\big(\text{Distances}_K(\mathbf{h}_i, \mathbf{E})\big).$$

Thus, low $\mathbf{H}_i$ implies high "copy" availability.

Predictive Power ($\mathbf{I}_i$) as "Relations". The "quantity of relations" is interpreted as the token's predictability and informational contribution. Following the original principle, we define predictive power as an inverse function of uncertainty:

$$\mathbf{I}_i = \frac{1}{\ln(\mathbf{H}_i + \epsilon) + 1}, \quad \epsilon = 10^{-6}.$$

A low-uncertainty token (low $\mathbf{H}_i$) yields high $\mathbf{I}_i$, reflecting strong semantic integration (many "relations").

This balance reflects an informational equilibrium in the token's representational role. If a token's hidden state implies more outgoing semantic dependencies (high "Relations") than incoming contextual support (low "Copies"), the model lacks sufficient grounding to resolve the correct generative path—manifesting as attentional defocusing or hallucination. Conversely, if contextual support vastly exceeds outgoing dependencies (excess "Copies"), the representation becomes overly constrained, leading to information compression or repetition—a form of representational "looping". The $\varphi_i \approx 1.0$ condition thus corresponds to a state of optimal information flow, where context and prediction are in dynamic equilibrium.

Final Metric and Loss. The adapted φ-metric preserves the original ratio and target:

$$\varphi_i = \frac{\mathbf{I}_i}{\mathbf{H}_i}.$$

We apply a quadratic penalty to drive φ_i toward 1.0:

$$\mathcal{L}_{\varphi,i} = (\varphi_i - 1.0)^2.$$

The total loss combines this with the standard negative log-likelihood (NLL):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \lambda \cdot \frac{\text{Mean}(\mathcal{L}_{\varphi,i})}{S},$$

where λ is the regularization strength (LAMBDA) and $S = 10^4$ is a scaling factor.

We fine-tune the TinyLlama-1.1B model [9] using QLoRA [2] with 4-bit NF4 quantization [3]. Adaptation is applied to attention projections via LoRA [4] with rank $r = 8$, $\alpha = 16$. Training uses the `english_quotes` dataset [**?**] and is implemented using Hugging Face Transformers and PEFT [8, 5]. Nearest-neighbor search leverages FAISS [6].

Fifteen diverse prompts were used:

- Scientific explanation (photosynthesis)
- Historical timeline (ARPANET → WWW)
- Philosophical comparison (Dewey vs. Kant, Stoicism vs. Epicureanism)
- Strategic frameworks (SWOT, investment analysis)
- Creative/abstract reasoning (emergence, saudade, insight as geometry)

We propose a novel regularization term that penalizes deviation of the per-token φ-metric from the optimal value of 1.0. The φ-metric balances predictive power and semantic uncertainty by comparing each token's hidden state to its nearest neighbors in the embedding space. The full training procedure is summarized in Algorithm 1.

4 Results

Table 1: Qualitative performance across LAMBDA configurations

Criterion	**LAMBDA = 0.0**	**LAMBDA = 0.1**	**LAMBDA = 0.25**	**LAMBDA = 0.5**
Completion	Poor	Poor	Moderate	Moderate
Factual Accuracy	Poor	Poor	Good	Moderate
Logical Coherence	Poor	Moderate	Good	Good
Repetition / Looping	Poor	Moderate	Good	Good
Prompt Adherence	Moderate	Moderate	Good	Good
Response Depth	Poor	Moderate	Good	Moderate
Overall Reliability	Poor	Moderate	**Good**	Moderate

Algorithm 1 φ-Regularized Training Step

Require: Input tokens $\mathbf{x} \in \mathbb{V}^L$, embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathbb{V}| \times d}$,
FAISS index $\mathcal{I}$ built over $\mathbf{E}$, regularization strength λ, scale $S = 10^4$
Ensure: Total loss $\mathcal{L}_{\text{total}}$

Forward pass: obtain hidden states $\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(L)} = \text{Model}(\mathbf{x})$
Use last-layer hidden states: $\mathbf{h}_i \leftarrow \mathbf{H}_i^{(L)}$ for all i
(Optional) Reduce dimensionality: $\mathbf{h}_i \leftarrow \mathbf{h}_i[:d_{\text{sub}}]$, where $d_{\text{sub}} = 256$
Identify active tokens: $\mathcal{T}_{\text{active}} = \{i \mid x_i \neq \texttt{[PAD]}\}$
for each $i \in \mathcal{T}_{\text{active}}$ **do**
  Retrieve $K = 5$ nearest neighbor embeddings:
  $\mathbf{E}_{\text{NN}_K} \leftarrow \text{TopK}(\mathcal{I}.\text{search}(\mathbf{h}_i), K)$
  Compute semantic uncertainty:
  $\mathbf{H}_i \leftarrow \frac{1}{K} \sum_{j=1}^{K} \|\mathbf{h}_i - \mathbf{E}_{\text{NN}_K}^{(j)}\|_2$
  Compute predictive power:
  $\mathbf{I}_i \leftarrow \frac{1}{\ln(\mathbf{H}_i + \epsilon) + 1}, \quad \epsilon = 10^{-6}$
  Compute φ-metric:
  $\varphi_i \leftarrow \mathbf{I}_i / \mathbf{H}_i$
end for
Compute NLL loss: $\mathcal{L}_{\text{NLL}}$
Compute φ-regularization loss:
$\mathcal{L}_\varphi \leftarrow \frac{1}{|\mathcal{T}_{\text{active}}|} \sum_{i \in \mathcal{T}_{\text{active}}} (\varphi_i - 1.0)^2$
Scale and combine:
$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{NLL}} + \lambda \cdot \frac{\mathcal{L}_\varphi}{S}$
return $\mathcal{L}_{\text{total}}$

4.1 Key Observations

- **LAMBDA = 0.0 (Baseline)**: Frequent factual errors (e.g., "plants emit light", "TCP/IP launched in 1983", "Kant's CI in *Critique of Pure Reason*"), mid-sentence cuts (7/15 prompts), and repetitive phrasing.
- **LAMBDA = 0.1**: Reduction in hallucinations, but persistent timeline errors (e.g., "Gopher by Berners-Lee") and incomplete structures.
- **LAMBDA = 0.25**: Best balance—high accuracy, minimal repetition, strong prompt adherence (e.g., complete SWOT, correct TOC steps, nuanced philosophy). Occasional cuts due to token limits.
- **LAMBDA = 0.5**: Structurally sound but prone to oversimplification (e.g., "Epicureans = pursuit of food/drink", "Kant = quantitative").

5 Conclusion

φ-regularization significantly improves model reliability. The optimal strength is **LAMBDA = 0.25**, achieving the best trade-off between factual correctness, coherence, and task fidelity. This supports the hypothesis that internal representation stability—guided by $\varphi_i \to 1.0$—enhances generative quality.

References

[1] Anonymous. Mechanics of mind. https://archive.org/details/20251128_20251128_0859, 2025. Accessed: 2025-12-05.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized large language models. *arXiv preprint arXiv:2305.14314*, 2023.

[3] Tim Dettmers and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Hugging Face. Peft: Parameter-efficient fine-tuning. https://github.com/huggingface/peft, 2023. Accessed: 2025-12-05.

[6] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Thibaut Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[9] Zijun Zhang, Tianle Wang, Zhihao Ding, Xuxiang Ma, et al. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Appendix: LAMBDA = 0.0 — Full Prompt-Response Comparison Across Decoding Strategies

All responses below are generated by the same underlying model (TinyLlama-1.1B with LAMBDA = 0.0), but under different decoding configurations. Only Prompts 1–3 were tested under all four settings; Prompts 4–15 are shown for the baseline configuration only.

Prompt 1: Explain the process of photosynthesis in three main steps, starting with light absorption.

Baseline (TEMP = 0.7): Sure, I'd be happy to explain the process of photosynthesis in three main steps, starting with light absorption:
1. Light Absorption: Light is the primary source of energy for photosynthesis. As plants and other organisms emit light, it reaches the surface of the leaves. There, the light energy is absorbed by chlorophyll, the green pigment in plant cells.
Chlorophyll absorbs light in three different ways:
1. Photons: These are the individual light particles that make up the electromagnetic radiation that plants and other organisms use for light absorption.
2. Pigments: The green pigment

(Truncated; contains critical error: "plants emit light")

Test 1 (max_tokens, T=1.0, top_p=1.0): Sure! Here's a breakdown of the process of photosynthesis in three main steps, starting with light absorption:
1. Light Absorption: Light energy is absorbed by chlorophyll molecules in the leaves of plants. Chlorophyll is a pigment that absorbs light in the visible spectrum, which is the range of light that we can see with our eyes. The absorbed light energy is used to drive the chemical reactions that occur in the chloroplasts.
2. Chlorophyll Reaction: The absorbed light energy is used to drive the chemical reactions that occur in the chloroplasts. These reactions involve the transfer of electrons from the electron transport chain to the pigments, which in turn releases energy in the form of ATP and NADPH.
3. NADPH and ATP Production: The energy released from the chlorophyll reaction is used to drive the synthesis of NADPH

(Truncated; incorrect electron flow description)

Test 2 (T=0.5, top_p=0.8): Sure! Here's a breakdown of the process of photosynthesis in three main steps:
1. Light Absorption: Light is the primary source of energy for photosynthesis. When light enters the chloroplasts, it is absorbed by the pigments chlorophyll a and chlorophyll b...
2. Carbon Dioxide and Water: ...The energy is used to split water molecules into oxygen gas (O2) and hydrogen gas (H2) by the enzyme chlorophyll a and b.
3. Light Released: The energy released by the process of photosynthesis is released into the environment in the form of light.

(Complete but contains critical error: "energy released as light")

Test 3 (T=0.7, top_p=1.0, rep_pen=1.15): Sure! Here's a simple explanation for how photosynthesis works:
1. Light Absorption (Light-Dependent Reactions): The first step is called the light-dependent reactions... Chlorophyll absorbs light at a specific wavelength... and converts it into electrical energy...
At this point, the electrons are pulled out of their binding sites by ATP molecules...
In this second part of the process, the electrons travel through the membrane...
The process begins when the electrons encounter an acceptor pigment known as complex I...
From there, the electrons are added to the

(Truncated; misattributes "complex I" from mitochondrial respiration)

Prompt 2: Provide a historical timeline (4 events) of the development of the internet from ARPANET to the World Wide Web.

Baseline (TEMP = 0.7): 1. ARPANET... established in 1969...
2. The internet is launched in 1983 by DARPA...
3. The first internet domain name, ".com," is registered on October 24, 1985.
4. The World Wide Web (WWW) is developed in the mid-1990s by Tim Berners-Lee...
5.

(Erroneous: WWW developed in late 1980s–1991, not "mid-1990s"; extra "5." suggests confusion)

Test 1: 1. ARPANET (Advanced Research Projects Agency Network) - 1969...
2. TCP/IP (Transmission Control Protocol/Internet Protocol) - 1973...
3. World Wide Web (WWW) - 1989...
4. World Wide Web Consortium (W3C) - 1994...
The World Wide Web Consortium (W3

(Truncated; correct sequence but incomplete)

Test 2: 1. ARPANET... created to facilitate communication...
2. ARPANET Expansion - 1970...
3. ARPANET/X.X.X Network Expansion - 1971...
4. The World Wide Web - 1991...
The World Wide Web (WWW) was created by Tim Berners-Lee, a British

(Fictional expansions "1970", "1971"; misattributes NBS as ARPANET origin)

Test 3: *(Not provided in data)*

Prompt 3: Explain the fundamental differences between John Dewey's concept of experience and Immanuel Kant's categorical imperative.

Baseline (TEMP = 0.7): ...Kant's categorical imperative is based on the idea that we have a duty to act in accordance with the categorical imperatives set forth in his book, "The Critique of Pure Reason."

(Critical error: CI is in Groundwork, not Critique of Pure Reason)

Test 1: John Dewey's concept of experience is based on the idea that all knowledge is subjective...
Kant's categorical imperative... is a moral principle that requires us to act in accordance with our understanding of what is right...
Dewey's concept of experience is central to his philosophy of education...
The importance

(Truncated; surface-level but no gross error)

Test 2: John Dewey and Immanuel Kant are two prominent philosophers...
1. Dewey's concept of experience: ...based on the idea that all knowledge is subjective...
2. Kant's categorical imperative: ...a principle that requires us to act in accordance with what is necessary and right...
Dewey's concept of experience is based on the idea that our knowledge is shaped by our experiences, and that our

(Truncated; oversimplified but coherent)

Test 3: *(Not provided in data)*

Prompts 4–15: LAMBDA = 0.0 (Baseline, TEMP = 0.7 only)

Prompt 4 (Stoicism vs Epicureanism): Truncated after "Stoics believe in the concept of time as"

Prompt 5 (Eigenvalue): Truncated mid-formula: "R = (x12, x2"

Prompt 6 (Investment Framework): Truncated after point "4"

Prompt 7 (Team Conflict): Truncated mid-sentence: "3. Reframe the problem: Instead of simply addressing the root"

Prompt 8 (Theory of Constraints): Truncated after "c. Knowledge constraints: The system must maintain a"

Prompt 9 (AI Maturity Model): Only 3 of 5 stages provided; cuts off mid-sentence

Prompt 10 (SWOT): Truncated after "Opportun"

Prompt 11 (Geometrical Insight): Truncated after "but"

Prompt 12 (Emergence): Truncated after "1. The human mind is like a symphony orchestra..."

Prompt 13 (Minimalist Writer): Truncated after "Once I have identified the essential idea, I"

Prompt 14 (Saudade): Truncated after "3) Saudade is a feeling of longing for a time that"

Prompt 15 (Knowledge vs Understanding): Truncated mid-explanation

Summary: Even with advanced decoding strategies, the LAMBDA = 0.0 model consistently fails to complete responses and produces significant factual errors. Only Prompts 1–3 were tested under all decoding regimes; for the remaining prompts, only the baseline (TEMP = 0.7) outputs are available.

Appendix: Full Prompt–Response Comparison Across LAMBDA Values

All responses below are generated by the φ-regularized TinyLlama-1.1B model under identical prompt conditions, with only the regularization strength (λ) and temperature varied:

- LAMBDA = 0.0, TEMP = 0.7 (Baseline)
- LAMBDA = 0.1, TEMP = 0.6
- LAMBDA = 0.25, TEMP = 0.5
- LAMBDA = 0.5, TEMP = 0.5

Prompt 1: Photosynthesis

LAMBDA = 0.0: "As plants and other organisms **emit light**, it reaches the surface..." → *Critical error; truncated.*

LAMBDA = 0.1: "...chlorophyll a, which is present in the **stroma**..." → *Mislocation; truncated.*

LAMBDA = 0.25: "...converted into carbohydrates through a series of chemical reactions..." → *Oversimplified but coherent; truncated.*

LAMBDA = 0.5: "Chlorophyll **binds to carbon dioxide**..." → *Biologically incorrect; truncated.*

Prompt 2: Internet Timeline

LAMBDA = 0.0: "The internet is launched in 1983 by DARPA... WWW in **mid-1990s**..." → *Inaccurate chronology.*

LAMBDA = 0.1: "Gopher developed in 1988 by **Tim Berners-Lee**..." → *Factual error; truncated.*

LAMBDA = 0.25: Correct sequence: ARPANET → TCP/IP → WWW; *minor inaccuracy (Berkeley for TCP/IP); truncated.*

LAMBDA = 0.5: "ARPAnet... the **first internet**, created in 1983..." → *Redundant, vague; truncated.*

Prompt 3: Dewey vs. Kant

LAMBDA = 0.0: "...categorical imperative... in **The Critique of Pure Reason**." → *Wrong core work.*

LAMBDA = 0.1: "Kant believed knowledge is based on the categorical imperative..." → *Conceptual confusion.*

LAMBDA = 0.25: Nuanced contrast: experience as dynamic vs. static; *no gross errors; truncated.*

LAMBDA = 0.5: "Dewey = qualitative, Kant = **quantitative**..." → *Misrepresentation.*

Prompt 4: Stoicism vs. Epicureanism

LAMBDA = 0.0: Truncated after "Stoics believe in the concept of time as..."

LAMBDA = 0.1: Correct contrast; truncated mid-sentence.

LAMBDA = 0.25: Well-structured comparison of virtue vs. pleasure; *truncated but coherent.*

LAMBDA = 0.5: "Epicureanism = pursuit of pleasure through **food and drink**" → *Reductive caricature.*

Prompt 5: Eigenvalue

LAMBDA = 0.0: Formula cut off: "R = (x12, x2..."; vague definition.

LAMBDA = 0.1: Better definition but imprecise; truncated.

LAMBDA = 0.25: Mentions eigenvalue decomposition, SVD formula; technical but flawed dimensions.

LAMBDA = 0.5: "Eigenvalue = multiplier of a specific row/column" → *Incorrect; overly simplistic.*

Prompt 6: Investment Framework

LAMBDA = 0.0: Truncated after point 4.

LAMBDA = 0.1: Lists pros/cons but disorganized; truncated.

LAMBDA = 0.25: Clear pros/cons + goal alignment; truncated mid-section.

LAMBDA = 0.5: Structured and balanced; one of the few near-complete responses.

Prompt 7: Team Conflict

LAMBDA = 0.0: Truncated mid-sentence.

LAMBDA = 0.1: Addresses "two team members" (misreads prompt); truncated.

LAMBDA = 0.25: Empathetic, practical steps; truncated before conclusion.

LAMBDA = 0.5: Action-oriented but misses emotional nuance; truncated.

Prompt 8: Theory of Constraints

LAMBDA = 0.0: Lists 3 constraint types; truncated.

LAMBDA = 0.1: "1. Define the Problem" — not standard TOC; truncated.

LAMBDA = 0.25: Reasonable 3-step outline; truncated.

LAMBDA = 0.5: Adds "bottlenecks" — acceptable adaptation; truncated.

Prompt 9: AI Maturity Model

LAMBDA = 0.0: 3 stages only; incomplete.

LAMBDA = 0.1: 3 stages; governance mentioned; truncated.

LAMBDA = 0.25: All 5 stages clearly defined (governance, data infra, security); truncated in last point.

LAMBDA = 0.5: 4 stages with operational focus; truncated.

Prompt 10: SWOT Analysis

LAMBDA = 0.0: Truncated after "Opportun".

LAMBDA = 0.1: Full SWOT; generic but correct structure.

LAMBDA = 0.25: Detailed, specific strengths (brand, talent, market position); truncated in Threats.

LAMBDA = 0.5: Concise and practical; complete structure.

Prompts 11–15: Creative and Abstract Tasks

For all four models, responses are **consistently truncated** in Prompts 11–15. However:

- **LAMBDA = 0.0** often produces vague or circular language.
- **LAMBDA = 0.1–0.25** show greater fluency and thematic coherence.
- **LAMBDA = 0.5** occasionally breaks persona (e.g., "I am not a famous minimalist writer").

Overall Observation: Only LAMBDA = 0.25 consistently balances accuracy, structure, and adherence to prompt across scientific, historical, philosophical, and strategic domains — despite shared token-limit truncation.